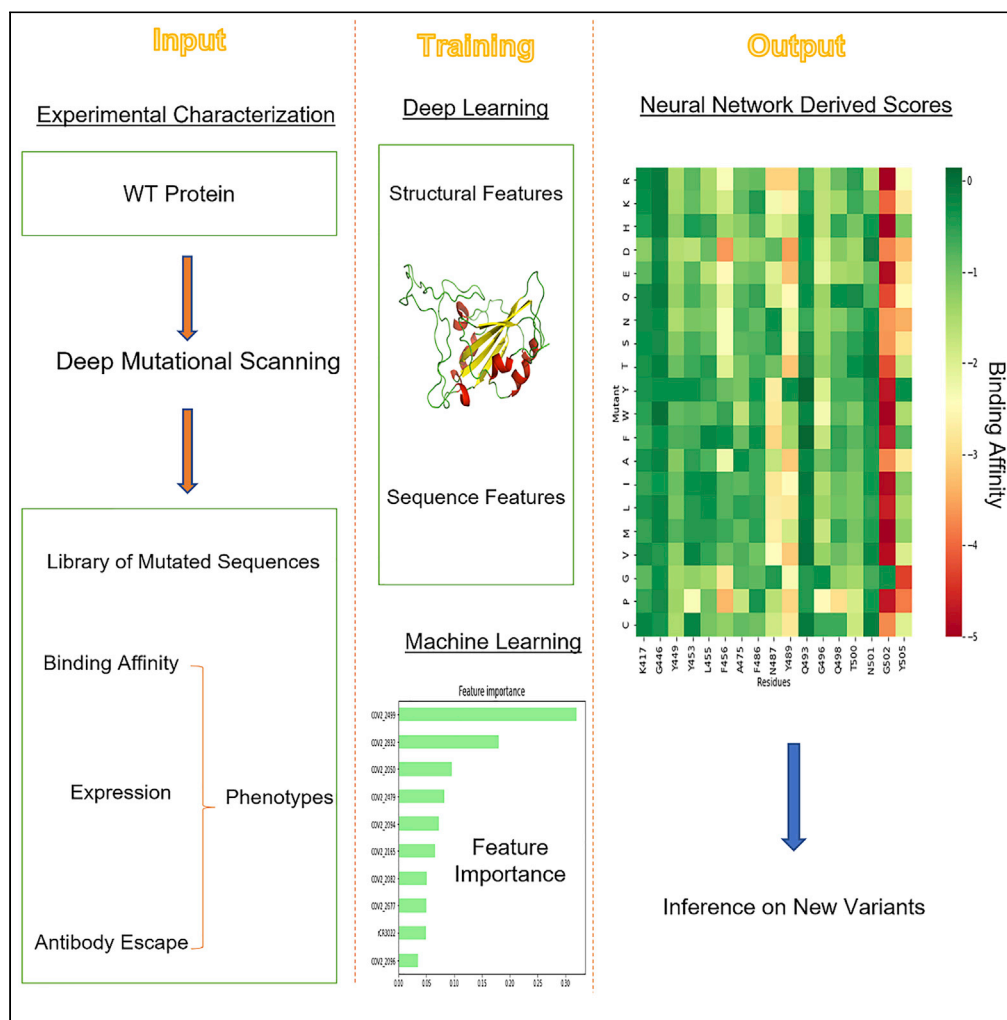


Article

# Modeling mutational effects on biochemical phenotypes using convolutional neural networks: application to SARS-CoV-2



Bo Wang, Eric R. Gamazon

eric.gamazon@vumc.org

**Highlights**

Deep learning models of biochemical phenotypes from deep mutational scanning (DMS) data

Prediction performance gain from using physicochemical properties of amino acids

Concordance of neural network predictions with molecular dynamics simulations

Improved causal inference properties for neural-network-defined phenotypes

Wang & Gamazon, iScience  
25, 104500  
July 15, 2022 © 2022 The Author(s).  
<https://doi.org/10.1016/j.isci.2022.104500>



## Article

## Modeling mutational effects on biochemical phenotypes using convolutional neural networks: application to SARS-CoV-2

Bo Wang<sup>1</sup> and Eric R. Gamazon<sup>1,2,3,4,5,\*</sup>

## SUMMARY

Deep mutational scanning (DMS) experiments have been performed on SARS-CoV-2's spike receptor-binding domain (RBD) and human angiotensin-converting enzyme 2 (ACE2) zinc-binding peptidase domain—both central players in viral infection and evolution and antibody evasion—quantifying how mutations impact biochemical phenotypes. We modeled biochemical phenotypes from massively parallel assays, using neural networks trained on protein sequence mutations in the virus and human host. Neural networks were significantly predictive of binding affinity, protein expression, and antibody escape, learning complex interactions and higher-order features that are difficult to capture with conventional methods from structural biology. Integrating the physicochemical properties of amino acids, such as hydrophobicity and long-range non-bonded energy per atom, significantly improved prediction (empirical  $p < 0.01$ ). We observed concordance of the neural network predictions with molecular dynamics (multiple 500 ns or 1  $\mu$ s all-atom) simulations of the spike protein-ACE2 interface, with critical implications for the use of deep learning to dissect molecular mechanisms.

## INTRODUCTION

Since the initial outbreak, the SARS-CoV-2 virus has rapidly spread worldwide causing a global public health crisis, the coronavirus disease 2019 (COVID-19). *In vitro* and cryo-electron microscopy studies have established that the betacoronavirus uses the human cell-surface protein angiotensin-converting enzyme 2 (ACE2) to gain entry into target cells (Lan et al., 2020; Zhang et al., 2020; Zhou et al., 2020). Therefore, precise characterization of the interaction between the receptor-binding domain (RBD) of the viral spike glycoprotein and the ACE2 complex is of critical importance in understanding COVID-19 pathophysiology (Zhou et al., 2020). Not surprisingly, several drug candidates that target either the virus or the receptor have been developed on the basis of the ACE2 binding. With an improved understanding of this key molecular interaction, two major therapeutic strategies have been pursued, including (1) engineering high-affinity ACE2 decoy or developing antibody cocktail treatments and (2) screening new or repurposing existing inhibitors targeting the binding interface (Chan et al., 2020; Han and Král, 2020). Establishing the sequence-structure-phenotype relationship for the spike RBD and the ACE2 receptor is essential for both strategies, in which the sequence mutational effect on receptor affinity and other biochemical phenotypes is the major component (Blanco et al., 2020; Calcagnile et al., 2021; Hussain et al., 2020; Procko, 2020; Starr et al., 2020).

Comprehensive understanding of how variants, from single mutations to their combinations, affect disease-relevant biochemical phenotypes would go a long way toward clarifying molecular mechanisms of disease as well as downstream adverse complications and guiding pharmacological interventions. In addition, elucidating the mutational effect may shed light on selective pressures determining the evolutionary trajectory of the coronavirus as well as identify risk factors for viral infection and maladaptive host response to COVID-19 in human populations (Pietzner et al., 2020). Deep mutational scanning (DMS) systematically evaluates the effect of mutant versions of the protein on measured biochemical phenotypes (Fowler et al., 2010; Fowler and Fields, 2014; Procko, 2020; Stein et al., 2019). High-throughput mutagenesis in DMS makes it possible to assess the phenotypic consequences of each possible amino acid mutation in a protein, generating large datasets that can reveal the sequence-function landscape. The development of computational approaches to learn the complex and non-linear features of this map can enable

<sup>1</sup>Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN 37232, USA

<sup>2</sup>Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN 37232, USA

<sup>3</sup>Data Science Institute, Vanderbilt University Medical Center, Nashville, TN 37232, USA

<sup>4</sup>Clare Hall, University of Cambridge, Cambridge CB3 9AL, UK

<sup>5</sup>Lead contact

\*Correspondence:

[eric.gamazon@vumc.org](mailto:eric.gamazon@vumc.org)

<https://doi.org/10.1016/j.isci.2022.104500>



high-throughput inference of basic protein properties. Statistical and machine learning methods, including deep learning, have attracted significant attention owing to their predictive power (Angermueller et al., 2016). In a recently developed supervised learning framework tailored to DMS datasets, convolutional neural networks demonstrated spectacular performance, consistent with other recent studies of mutational effect (Gelman et al., 2020; Li et al., 2020).

DMS experiments on both the SARS-CoV-2 spike glycoprotein and the ACE2 receptor have been performed, providing an important basis for further investigations of mutational effects (Chan et al., 2020; Heinzelman and Romero, 2021; Starr et al., 2020). In this work, we conducted systematic modeling of the mutational effects of the RBD in the viral spike protein and of the ACE2 receptor on biochemical phenotypes, extending a supervised learning framework (Gelman et al., 2020). Three classes of critical phenotypes—binding affinity, protein expression, and antibody escape—were systematically analyzed within the sequence-structure-phenotype paradigm that informs much of proteomic and structural biology studies. Neural networks were also leveraged to learn (a) the complex functional landscape of the viral spike protein's RBD and the host cell-surface receptor ACE2 and (b) the antibody-escape map of RBD mutations, including mutations that undergo selection during viral proliferation in the presence of antibodies and mutations that have been circulating in human populations (Elbe and Buckland-Merrett, 2017; Greaney et al., 2020; Shu and McCauley, 2017). Finally, the mutation- or genetically—determined phenotypes (i.e., the component of a phenotype defined by the sequence versus a technical confounder or the environmental component) estimated from the neural networks were exploited to extract key insights into the molecular mechanisms of COVID-19 infection and severity. The neural networks promise to substantially extend DMS studies and can be used to predict the effects of complex mutations (e.g., combinations of single-point mutations) for downstream experimental studies.

## RESULTS

### Overview of framework

Briefly, we describe the framework (Figure 1A). Here we consider a set of  $N$  training samples consisting of (protein) sequences,  $x_1, x_2, \dots, x_N \in X$ , with corresponding biochemical phenotype values,  $y_1, y_2, \dots, y_N \in Y$  (where the phenotype is fixed and chosen from binding affinity, protein expression, or an antibody escape measure), where the pairs are sampled independently and identically distributed (*i.i.d.*) from a joint distribution

$$(x_i, y_i) \in \Delta$$

A supervised learning method seeks to find, from a pre-defined class of functions, a function  $g: X \rightarrow Y$ , which maps the input  $x$  to the phenotype value  $y$  and minimizes the expected deviation

$$E[L(y, g(x))] = \int L(y, g(x)) \Delta(x, y) dx dy$$

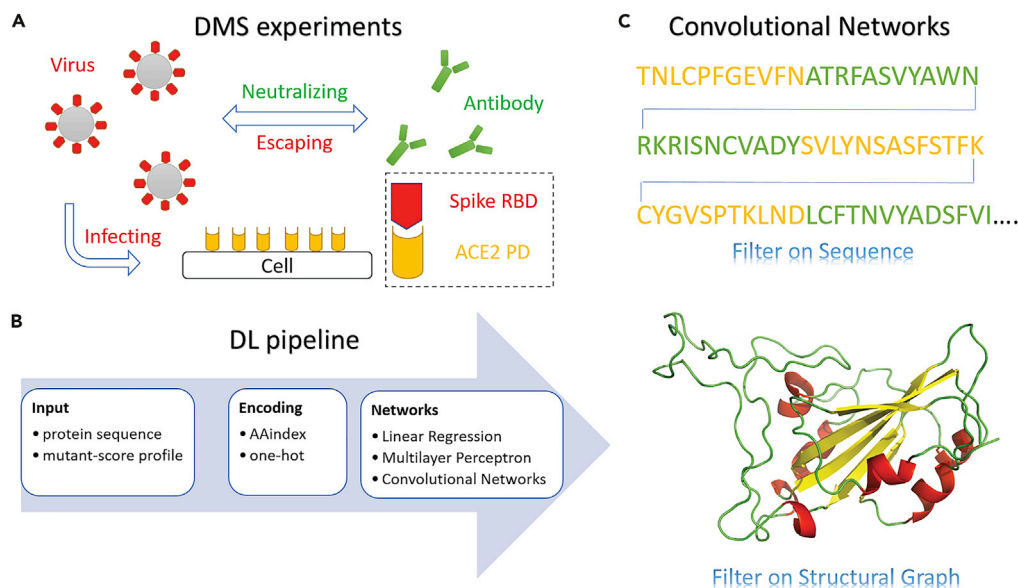
from  $y$  under a loss function  $L: X \times X \rightarrow Y$ . Regularization (e.g., dropout (Srivastava et al., 2014)) is also performed to prevent overfitting and enhance generalization. The expected loss is generally approximated by the average of the  $L(y_i, g(x_i))$  across the training samples, that is, the empirical loss, given that the distribution  $\Delta$  is typically not known. Neural networks constitute the class of functions defined by the composition of affine maps ( $\Psi_i$ ) and activation functions ( $\phi$ ) (which are typically non-linear).

$$g(x) = \Psi_M(\phi(\Psi_{M-1}(\phi(\dots\phi(\Psi_1(x)))))) \quad (\text{Equation 1})$$

(Note linear regression can be trivially included as a member of this class of functions.) The minimization problem, that is, the search for an optimal function  $g$ , is framed as a search for a set of weights  $\mathbf{W}_{(k)}$  and biases  $b_{(k)}$  (STAR Methods).

### Application of deep learning framework to the spike receptor-binding domain and the angiotensin-converting enzyme 2 peptidase domain complex

The crystal structure of the spike RBD—ACE2 peptidase domain (PD) complex was retrieved from the Protein Data Bank (PDB; PDB ID: 6M0J) (Lan et al., 2020). Our study investigated two biochemical phenotypes, binding affinity and protein expression—each critical to viral evolution, receptor recognition, and host-virus interaction—and a set of antibody-escape phenotypes, which may provide critical insights into the antigenic consequences of mutations and effective vaccine design (Figure 1A). Specifically, we trained on single-mutation functional score profiles, consisting of a large proportion of the amino acid mutations and the corresponding phenotypic measurements (quantified relative binding score for the ACE2 catalytic



**Figure 1. Overview of the framework**

(A) We evaluated three biochemical phenotypes generated from Deep Mutational Scanning (DMS) experiments: binding affinity, protein expression, and antibody escape. These biochemical phenotypes may determine the evolutionary trajectory of the coronavirus, viral infection, and maladaptive host response.

(B) We quantified the extent to which state-of-the-art neural network methodologies can predict these phenotypes, using sequence mutation data. The predicted biomedical phenotypes are mutation- or genetically determined traits, which are, therefore, of intrinsic interest because they are determined by the (amino acid) sequence data. We encoded the protein sequence input using AAindex, consisting of 566 intrinsic (data-independent) physicochemical properties of the amino acids, such as hydrophobicity and long-range non-bonded energy per atom, and one-hot encoding. We then trained neural networks on the mutation-phenotype data from the DMS experiments.

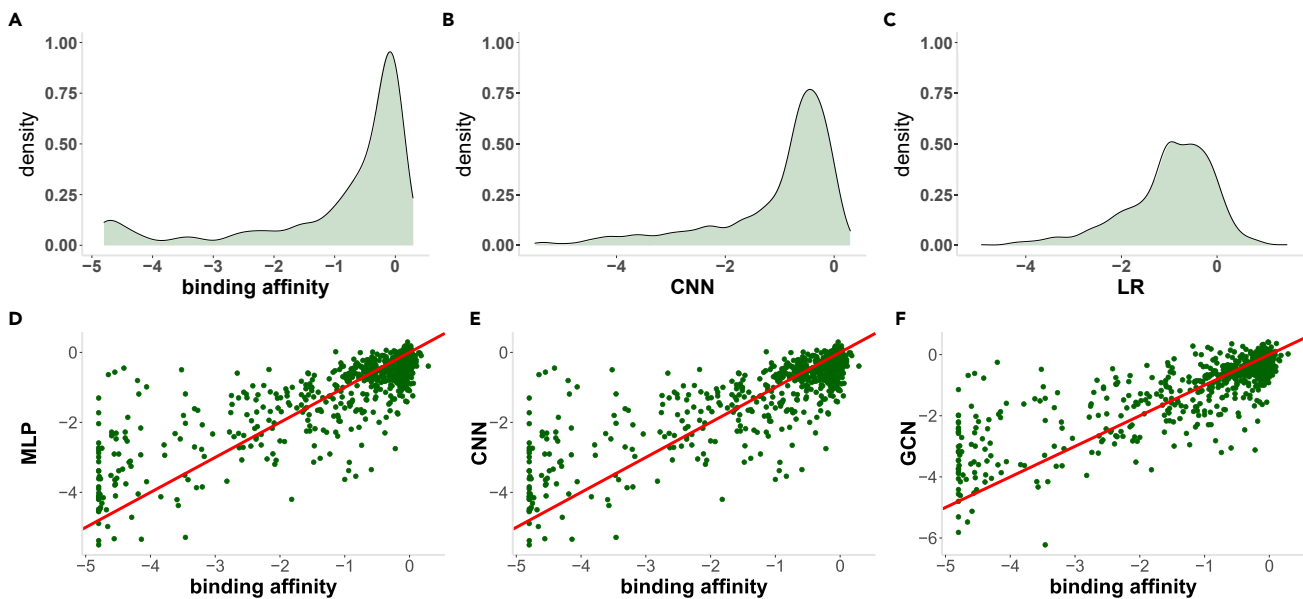
(C) We trained sequence convolutional neural networks and graph convolutional neural networks (along with [fully connected] multilayer perceptrons and linear regression) on the DMS data. Shown are the convoluted sequences and structural motifs, differentiated by color, for the two types of architecture, respectively.

zinc-binding PD, binding affinity, and expression for the spike RBD, and the “escape fraction” from each of 10 human monoclonal antibodies) measured by DMS experiments. The protein sequences were encoded by one-hot encoding and an external curated featurization table, AAindex, which summarizes the intrinsic physicochemical properties of amino acids, for example, hydrophobicity, solvent-accessible surface area, and long-range non-bonded energy per atom (Figure 1B; STAR Methods).

We generated independent models, which allow the prediction of mutational effects on binding affinity, relative binding score, expression, or antibody escape (for each of the 10 antibodies) of previously uncharacterized variants. We implemented linear regression and several neural network architectures: multilayer perceptron (MLP) (fully connected), sequence convolutional neural network (CNN), and graph convolutional neural network (GCN). Although linear regression considers only the additive contributions of sequence residues, multilayer perceptron can capture non-linear relationships between sequence (input) and phenotype (output). Owing to their distinctive architecture, convolutional neural networks are even more powerful to learn higher-level features. Filters for sequence convolutional neural networks slide along the one-dimensional input of protein sequence and learn from sequence fragments while graph convolutional neural networks act on the structural coordinates of proteins (Figure 1C). In the graph structure representation  $G = (V, E)$  of the spatial structure of the protein, the sequence residues are represented as nodes  $V$  while the interactions for pairs of residues (determined by a distance threshold) are represented as edges  $E$ . Graph convolutional neural networks learn segments of protein structures, that is, motifs, with the convolutional kernels capturing translational invariance.

### Evaluating models learned from deep mutational scanning data

We performed the supervised learning methods on the various spike RBD and ACE2 PD datasets, comprising mutant-score profiles of binding affinity, expression, and antibody escape. For each model



**Figure 2. Predictive performance of neural networks on binding affinity**

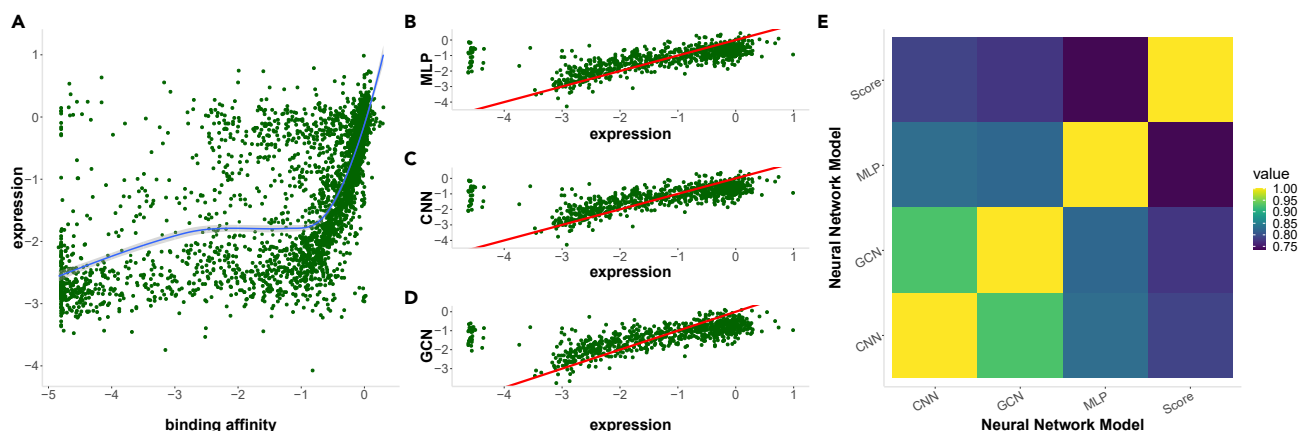
(A) Distribution of ACE2 binding affinity, in the test set, from single mutations in the spike glycoprotein RBD, as measured by DMS.

(B and C) The predicted biochemical phenotype from (B) neural networks, for example, CNN, was much closer in distribution to the original phenotype than that from (C) linear regression.

(D–F) Notably, (D–F) the three neural network models trained here showed a significant Spearman correlation ( $p < 2.2 \times 10^{-16}$ ) between the observed (x axis) and predicted (y axis) phenotype in the (independent) test set. Red line is the diagonal ( $y = x$ ) line. The binding affinity is from the original experimental study (ref 7), reported as  $\Delta \log_{10} K_{D,app}$ , where  $K_{D,app}$  is the “apparent dissociation constant” in molar units (M). Predicted binding affinity from a neural network is labeled as follows: CNN: convolutional neural network (Table S1, Network 2); LR: linear regression; MLP: multilayer perceptron; GCN: graph convolutional network (Table S1, Network 8).

(defined by a phenotype from a DMS dataset), the dataset was randomly split into 60% training, 20% tuning, and 20% testing subsets. Trained models were evaluated by predictive performance on the test subset (e.g., Figure 2A) with the Spearman correlation coefficient between the observed and predicted phenotype as a metric. The supervised learning methods demonstrated varied predictive performance in the case of the spike RBD, where the Spearman correlation ranged from 0.49 (for linear regression) to 0.78 (for convolutional neural network) for ACE2 binding affinity (Table S1). Neural networks provided much better predictive performance than linear regression, indicating that binding affinity is characterized by interactions (non-additive effects) among residues and a non-linear mapping of sequence to function. Relative to the distribution of binding affinity (Figure 2A) from the RBD mutations, the predicted phenotypes from the neural networks had greater fidelity to the original than linear regression (Figures 2B and 2C). Notably, the amino acid sequence mutational data alone could significantly predict binding affinity ( $p < 2.2 \times 10^{-16}$ ), using multilayer perceptron (Spearman correlation  $\rho = 0.52$ ), sequence convolutional neural network ( $\rho = 0.69$ ), and graph convolutional neural network ( $\rho = 0.70$ ) (Figures 2D–2F, Table S1).

For the spike glycoprotein expression, the prediction performance (Spearman correlation in the test set) ranged from 0.44 (for linear regression) to 0.84 (for graph convolutional network). Binding affinity to ACE2 and spike glycoprotein expression were significantly correlated ( $\rho = 0.65$ ) in the training set, but the relationship was found to be highly non-linear (Figure 3A). Though both phenotypes reflect protein stability under a wide range of biochemical contexts, binding affinity is more determined by the surface residues on or near the binding sites as well as by allosteric effects induced by inner or remote residues. As in the case of binding affinity, the neural network models significantly predicted RBD expression ( $p < 2.2 \times 10^{-16}$ ), using multilayer perceptron (Spearman correlation  $\rho = 0.50$ ), sequence convolutional neural network ( $\rho = 0.682$ ), and graph convolutional neural network ( $\rho = 0.679$ ) (Figures 3B–3D, Table S1). Despite the range in prediction performance of the models, the estimates of the mutation-mediated component of RBD expression derived from them were highly correlated (Figure 3E).



**Figure 3. Predictive performance of neural networks on protein expression**

(A) Binding affinity to ACE2 and spike glycoprotein expression, both induced by a protein sequence mutation on the RBD, were significantly correlated (Spearman correlation  $\rho = 0.65$ ), but the relationship between the two phenotypes was highly non-linear. We fit a curve through the points, using Locally Estimated Scatterplot Smoothing (LOESS), a method for local polynomial regression. The binding affinity is from the original experimental study (ref 7), reported as  $\Delta \log_{10} K_{D,app}$ , where  $K_{D,app}$  is the “apparent dissociation constant” in molar units (M). Expression measurement (ref 7) is the difference in log-mean fluorescence intensity (MFI) between variant and wild-type.

(B–D) We trained three neural network models, which could significantly predict (Spearman correlation  $p < 2.2 \times 10^{-16}$ ) protein expression in the (independent) test set. Red line is the diagonal ( $y = x$ ) line. X axis is observed expression.

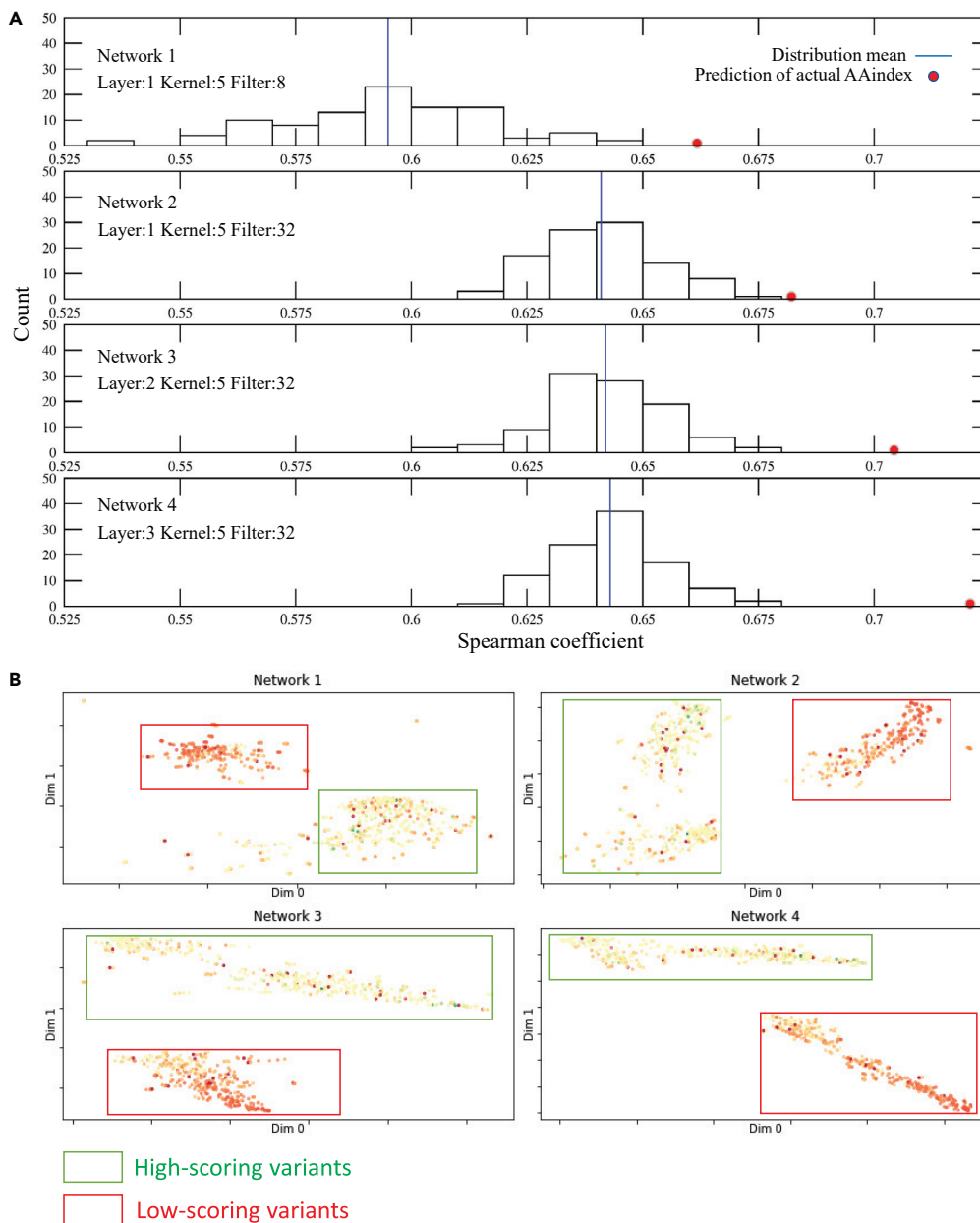
(E) Pairwise correlation between the neural network models. Predicted expression (y axis in B–D and x-, y axis in E) from a neural network is labeled as follows: MLP: multilayer perceptron; CNN: convolutional neural network (Table S1, Network 2); GCN: graph convolutional network (Table S1, Network 8).

Previous studies have shown that mechanistic insight of protein mutational effects is not easily transferable to studying the effects of variants in other proteins (Stein et al., 2019). We hypothesized that models trained on one species phenotype are not suitable for predicting another species phenotype. To evaluate this hypothesis, we trained a model using the avGFP DMS dataset (Sarkisyan et al., 2016), in which the score of the green fluorescent protein from *Aequorea victoria*, avGFP, is fluorescence. The avGFP sequence was slightly modified to match the input dimension of the spike RBD (STAR Methods). Therefore, the expression-based score profile of the spike RBD could be used as an external test set, with its combination of amino acid sequence unseen during the training and an unrelated phenotype. As hypothesized, the sequence convolutional model from avGFP failed to predict the expression of mutated sequences (Spearman correlation  $\rho = -0.0346$ ). This finding implies that the biochemical phenotypes differed in their dependence on protein amino acid sequence length, type, or composition, with interaction from non-sequence physicochemical environments (for example, pH condition and ionic density) potentially accounting for the difference.

### Influence of network architectures on biochemical phenotype prediction

All mutational sequences were encoded by the externally defined AAindex, a set of 566 quantitative indices of physicochemical properties of amino acids, which were used to train the empirical mutational scores on the DMS data (Gray et al., 2018). Thus, we examined the impact of the quality of the AAindex on prediction performance. We shuffled the entire AAindex so that the intrinsic information for each amino acid would be completely lost and replaced with random noise (STAR Methods). We generated 100 pseudo AAindex tables, on which sequence neural networks were trained, using a given choice of configurations, that is, number of layers and filter dimensions (Figure 4A). For each choice of network configurations, the distribution followed a bell-shaped curve, with the prediction performance from the actual AAindex located at the far-right and outperforming any of the shuffled tables. In other words, integrating the AAindex information significantly improved (empirical  $p < 0.01$ ) the overall performance of the neural networks.

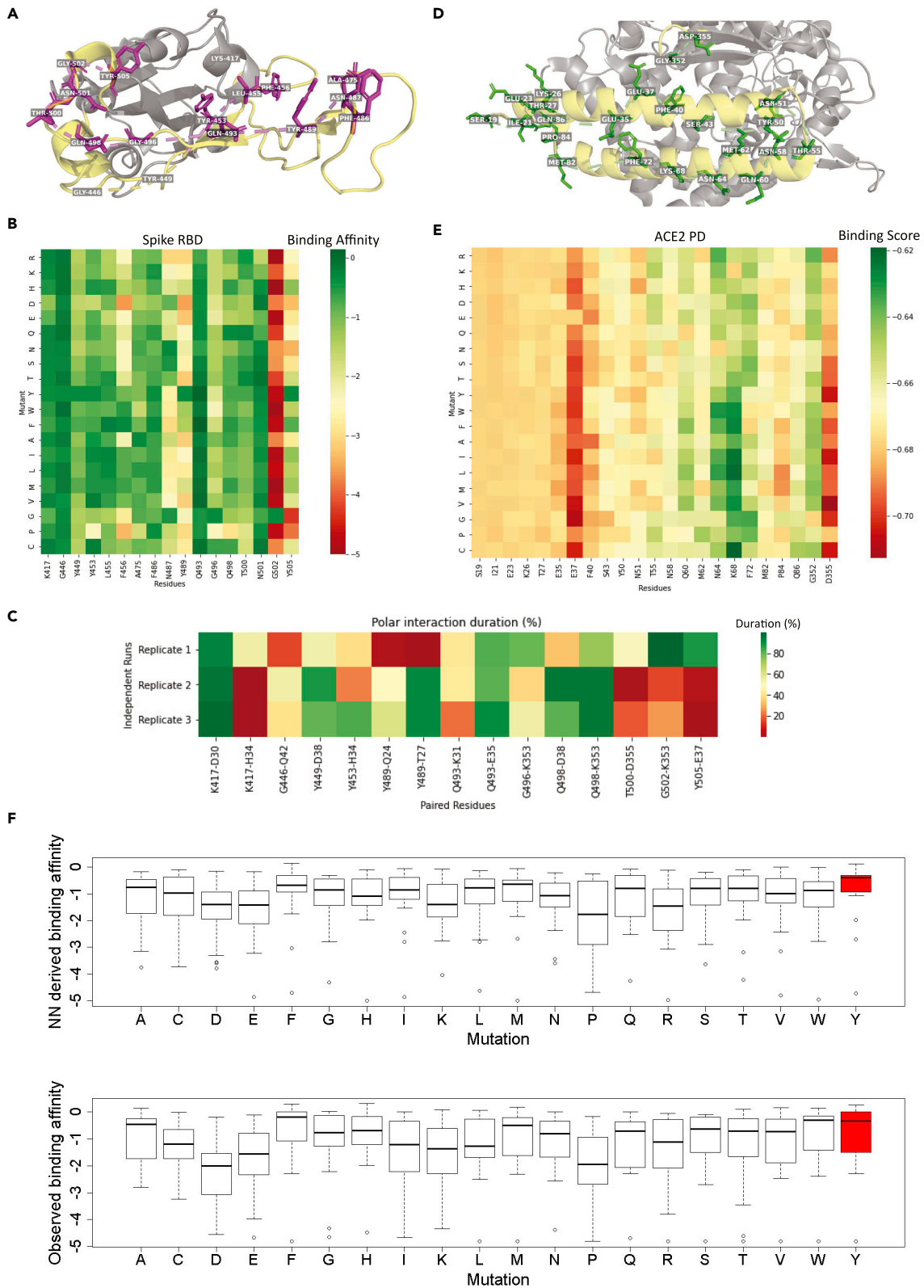
The prediction substantially improved with a larger filter dimension on a single hidden layer sequence convolutional network. This observation implies that the neural networks were able to learn more information from longer sequence fragments, which is consistent with the expression of the spike RBD reflecting complex non-linear relationships of the sequence (Network 1, 2 Figure 4A). Even with a shuffled AAindex table, that is, in which the intrinsic properties of the amino acids were essentially lost, a larger filter dimension led to enhanced learning from the expression phenotype, with increased mean and reduced SD for the



**Figure 4. Effect of neural network architectures on predicting biochemical phenotypes**

(A) Distribution of predictive performance (Spearman correlation in the test set) of sequence convolutional neural networks with a shuffled AAindex (shown as a histogram) and the actual AAindex (shown as a red dot), training on the spike RBD expression dataset. Each plot corresponds to a specific choice of architecture, including the number of layers and the filter dimension. Increasing the filter dimension or the number of layers improved the prediction performance. Furthermore, encoding the input data with the actual AAindex, which summarizes the externally defined intrinsic physicochemical properties of the amino acids, consistently led to performance gain. Nevertheless, we observed a strong dependence, on the input sequence data alone (i.e., without the intrinsic amino acid attributes), of the original biochemical phenotype, with reasonable performance across all choices of network configurations.

(B) Uniform Manifold Approximation and Projection (UMAP) projection of the latent space of the sequence convolutional networks on the last internal layer of the network. Mutants are colored according to the actual phenotype score, with green and red representing high-scoring and low-scoring variants, respectively.





**Figure 5. Mutational effects at key interface residues on the spike RBD and ACE2 PD**

(A and B) (A) Human ACE2-interacting residues (magenta) of the spike RBD interface (yellow) with (B) their predicted mutational effects on receptor affinity from a convolutional neural network with optimal performance (Network 3, Table S1). These key binding residues significantly differed (Mann-Whitney U test  $p = 5.19 \times 10^{-18}$ ) in the neural network-derived ACE2 binding affinity from the complement set of residues in the spike RBD, with the difference between the two residue classes more significant for this refined phenotype than for the original phenotype, that is, the phenotype unadjusted for non-mutational effects ( $p = 6.9 \times 10^{-14}$ ). Note the residues with outlier predicted mutational effects on binding affinity.

(C) Paired residues between the spike RBD and the ACE2 PD that retained polar interactions and the interaction duration percentage from molecular dynamics simulation trajectories. Briefly, 500 ns molecular dynamics simulations had been performed in triplicates using Desmond (Ali and Vijayan, 2020), with the structure of the spike RBD bound to ACE2 solvated with single point charge water molecules and with the use of the OPLS forcefield. Intermolecular interactions (e.g., hydrophobic interactions, hydrogen bonds,  $\pi$ - $\pi$ , cation- $\pi$ ) were dynamically observed in formation and breakage. We leveraged the residues of the spike RBD forming consistent polar interactions with the corresponding residues of ACE2 to evaluate concordance with the DMS measurements and our neural network predictions. Each paired residues included here were in polar interaction for at least 40% of the simulation duration in at least one replicate.

(D and E) (D) Human polymorphisms (shown in green stick representation), based on the gnomAD resource (v2) from large-scale sequencing projects, of the ACE2 PD interface with (E) the corresponding residues' predicted mutational effects on binding affinity from a convolutional neural network with optimal performance (Network 5, Table S2). Note the residues with outlier predicted mutational effects on the relative binding score.

(F) Among the key binding residues, a mutation to a Tyrosine (red) showed higher neural network-derived ACE2 receptor binding affinity and lower variability compared with a mutation to another amino acid. For comparison, a similar plot for the original phenotype (i.e., unadjusted for non-mutational effects) is shown in the bottom panel.

empirical performance distribution, indicating a highly strong dependence of expression on the sequence data. Deeper neural networks with a fixed filter dimension also showed improved prediction of expression (Network 2–4, Figure 4A). Integration of an accurate AAindex as input encoding of a deeper neural network resulted in further performance gain relative to a shuffled AAindex, confirming that the neural network was leveraging the externally defined intrinsic properties. To further reveal the information learned from neural networks by increasing the number of hidden layers, we applied Uniform Manifold Approximation and Projection (UMAP) to visualize the latent space in the last layer (Gelman et al., 2020; McInnes et al., 2020). Notably, a deeper network could better differentiate high- and low-scoring sequence mutants as distinct clusters (Network 1–4, Figure 4B).

**Mutational effects on biochemical phenotypes**

SARS-CoV-2 infection leads to complex pathophysiological consequences which are far from being fully understood. Elucidating mutational effects on biochemical phenotypes for the spike RBD and the ACE2 PD is critical in improving our understanding of the virus-host interactions as well as host immune response. Furthermore, with the spike glycoprotein as the primary target owing to its role in mediating coronavirus entry into host cells, neutralizing Antibodies (Abs) may provide protection from SARS-CoV-2 in humans (Addetia et al., 2020). Here, we highlight the results from sequence convolutional neural networks that showed the optimal performance among all tested models and further interpret them in connection with existing experimental characterizations, molecular modeling, and host genetic inferences. In addition, we leveraged the neural network-derived phenotypes to evaluate their utility in downstream analyses because, as inferred phenotypes, they are solely determined by the sequence mutation data and are thus less biased by possible confounding artifacts (Leek and Storey, 2007; Listgarten et al., 2010), including batch effects, environmental (non-sequence physicochemical) determinants, and other technical sources of heterogeneity.

*Key mutations on spike receptor-binding domain*

SARS-CoV-2 evolution promotes virulence and enhances escape from neutralizing Abs. Interface mutations (Figures 5A and 5D), specifically those at contact sites, are more likely to have a direct impact on infectivity. The mutational effects of these “key binding residues” of the spike RBD, that is, residues that are essential for ACE2 binding based on structural characterization (Lan et al., 2020), were systematically evaluated by our neural networks (Figures 5B and 5E). Predicted phenotypic effects were consistent with DMS measurements (Starr et al., 2020); for example, mutations at G502 showed the most extreme impact on receptor binding affinity (see Figure 5B). The neural network-derived (i.e., estimated, mutation-mediated) binding affinity with the ACE2 receptor significantly differed (Mann-Whitney U test  $p = 5.19 \times 10^{-18}$ ) between the set of key binding residues and the remaining residues in the spike RBD, thus inducing clustering. Notably, the difference between the two residue classes was more significant for this refined phenotype than for the original phenotype, that is, the binding affinity phenotype unadjusted for non-mutational effects ( $p = 6.9 \times 10^{-14}$ ). Interestingly, across the key binding residues in the RBD, a mutation to a Tyrosine showed higher neural network-derived receptor binding affinity and lower variability compared with a

mutation to another amino acid (Figure 5F). In addition, the key spike RBD binding residues were consistent with the results of interfacial interactions of the spike RBD bound to ACE2 from molecular dynamics simulations (Ali and Vijayan, 2020) (performed for 500 ns in triplicates or 1  $\mu$ s all-atom, using Desmond). Intermolecular interactions, including hydrophobic interactions, hydrogen bonds,  $\pi$ - $\pi$ , and cation- $\pi$ , were dynamically observed in formation and breakage in the simulations. Notably, the key binding residues of the spike RBD formed consistent polar interactions with the corresponding residues of ACE2 during the molecular dynamics simulation trajectories (Figure 5C).

Neutralizing Abs secreted from immune cells may also be found to be in complex with the spike RBD. Consistent with the experimental characterizations (Greaney et al., 2020), the neural network-derived antibody-escape phenotypes for the 10 human monoclonal Abs have differentiated target “hotspots” on the surface of the spike RBD, pointing to different sets of escape mutations (Figure 6A). The neural network prediction performance on antibody escape of the spike RBD varied between the Abs (Spearman correlation coefficient in the test set between 0.09 and 0.32) (Figure 6B). The variability in performance might be due, at least in part, to the unique physicochemical properties of the 10 monoclonal antibodies as well as their specific relative dependence on sequence mutation and other (non-sequence) determinants.

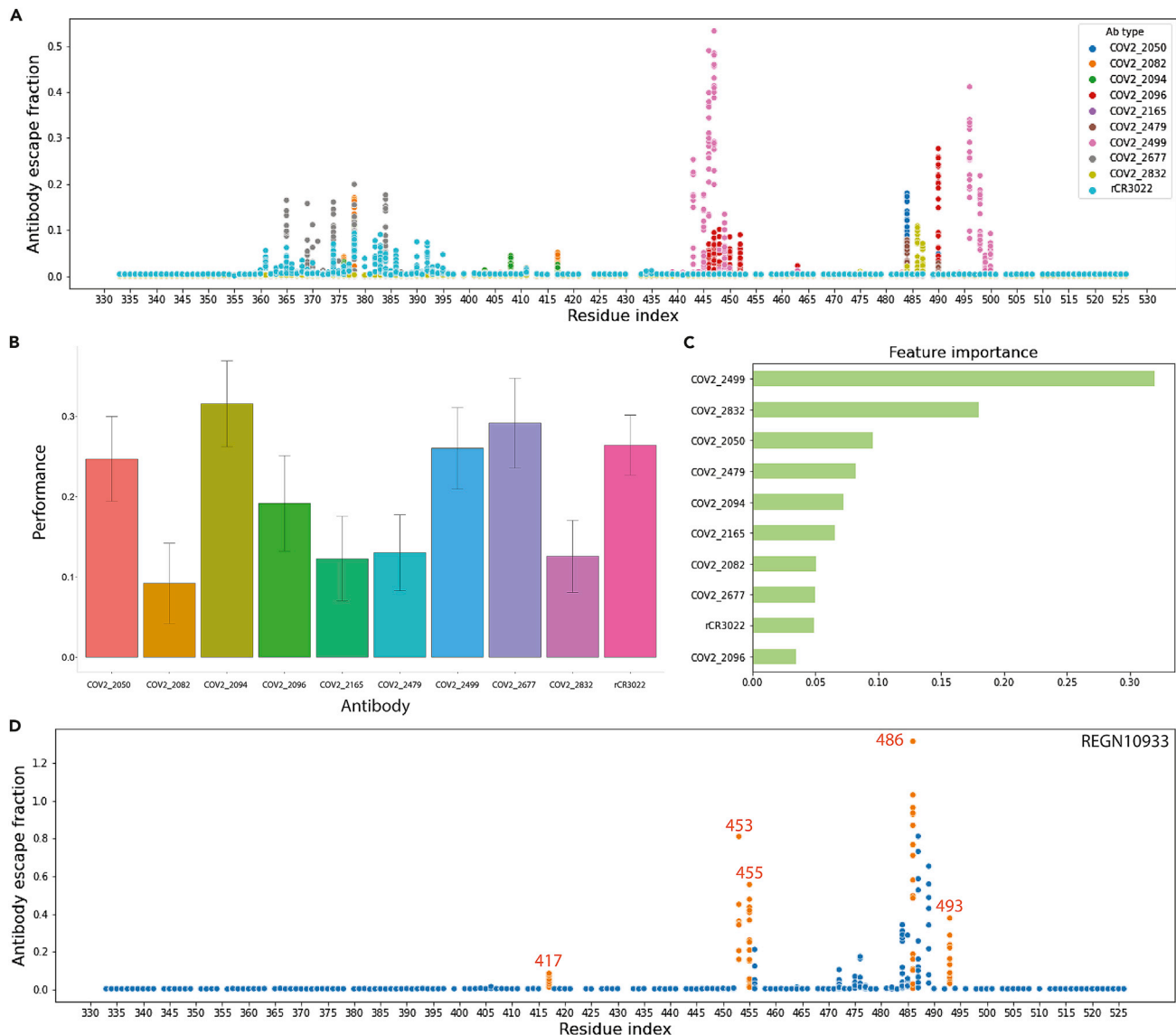
As receptor recognition and antibody escape are closely related to viral activities, we implemented Random Forest to calculate the *feature importance* of the 10 convolutional neural network-derived mutation-mediated antibody-escape phenotypes as predictors of binding affinity toward the ACE2 receptor. We hypothesized that those Abs that shield the receptor by pre-occupying the spike RBD—ACE2 PD binding interface would weigh more in their contributions to the prediction. Notably, we found that the top three Abs (COV2-2499, COV2-2832, and COV2-2050), all entirely targeting the interface based on experimental characterizations (Greaney et al., 2020), among the neural network-derived antibody-escape predictors accounted for nearly 60% of the feature importance score (Figure 6C). Key residues on the spike RBD that are critical binding sites with ACE2 PD include G446, Y449, G496, Q498, and T500 for escape mutations of COV2-2499. Both receptor affinity and the amount of available protein determine the infectivity of SARS-CoV-2 and downstream complications of COVID-19. Notably, among the key binding residues on the spike RBD, these sites of escape mutations for the antibody differed significantly (Mann-Whitney U  $p = 3.2 \times 10^{-4}$ ) in their neural network-derived binding affinity than the complement set of residues and, even more significantly ( $p = 1.03 \times 10^{-19}$ ), in their neural network-derived protein expression profile. Thus, the antibody escape mutations belonged to a sub-cluster within the key binding residue cluster (described above). In addition, we found these residues to be the sites of the peaks for the neural network-derived antibody-escape phenotype for COV2-2499, that is, the predicted escape mutations (Figure 6A). The importance of these residues for viral binding to the host receptor was also supported by results on intermolecular interactions from the molecular dynamics simulations (Figure 5C).

#### Key mutations on angiotensin-converting enzyme 2 peptidase domain

We comprehensively evaluated mutational effects on the relative binding of ACE2 PD. The corresponding key binding residues in the human ACE2 receptor paired with those of the spike RBD (Figure 5C) differed significantly (Mann-Whitney U  $p = 2.39 \times 10^{-7}$ ) in the neural network-derived (i.e., estimated, mutation-mediated) relative binding score from the remaining residues, with the key binding residues showing substantially lower variance (i.e., 40% of the complement set of residues). In addition, we investigated the subset of polymorphic sites in human populations and, using the neural network-derived phenotype, identified residues of outlier mutational effects on the relative binding score (Figures 5D and 5E). Although linear regression (Spearman correlation  $\rho = -0.014$ ) and the multilayer perceptron (Spearman correlation  $\rho = -0.044$ ) failed to make accurate predictions, the best performance achieved from sequence convolutional neural networks was 0.37 (Table S2). The lower phenotypic variance explained by the neural network models on ACE2 PD binding implies a complex sequence-function relationship, with possible contributions from sequence mutations and other (environmental) determinants. Notably, the empirical distribution of the score of the spike RBD and the ACE2 PD (Figure S1) is consistent with the spike RBD displaying more non-conserved mutations than the ACE2 PD, which fits our intuitive understanding of the properties of the virus (which must undergo adaptation to transmission and replication in the host) and the membrane receptor (with its conserved function).

#### Independent confirmation of neural network findings

We performed further confirmation of our specific neural network predictions and our overall methodology by analyzing an external dataset on an anti-SARS-CoV-2 spike monoclonal antibody REGN10933, which



**Figure 6. Neural network-derived antibody-escape phenotypes, performance prediction, and feature importance**

(A) Maps of the convolutional neural network-derived (i.e., estimated, mutation-mediated) escape fraction phenotype on the spike RBD for the different Abs (shown as different colors). Shifting peaks in different colors indicate differentiated binding sites and highlight different sets of escape mutations for the Abs on the spike RBD surface. Key residues on the spike RBD that are critical binding sites with ACE2 PD include G446, Y449, G496, Q498, and T500 for escape mutations of COV2-2499. Note these residues are the sites of the peaks of the neural network-derived antibody-escape phenotype for COV2-2499, that is, the predicted escape mutations. The importance of these residues for viral binding to the host receptor is also supported by results on intermolecular interactions from the molecular dynamics simulations (Figure 5C).

(B) Predictive performance (Spearman correlation in the test set between the observed and predicted antibody-escape fraction) of our neural network model on each of the 10 Abs (Table S3). For each antibody, the SE is shown, as generated from the application of *bootstrap* ( $n = 100$ ).

(C) Feature importance of the convolutional neural network-derived (i.e., estimated, mutation-mediated) antibody-escape phenotype for each of the 10 Abs as derived from a joint Random Forest regression model in predicting binding affinity toward the ACE2 receptor. The Abs that shield the receptor by entirely targeting the binding interface ranked in the top three neural network-derived antibody-escape predictors.

(D) Map of the neural network-derived (i.e., estimated, mutation-mediated) antibody-escape phenotype for the monoclonal antibody REGN10933. Here, we show the convolutional neural network model (Network 6) with optimal performance. Note the several peaks, that is, escape mutations, at the residues K417, Y453, L455, F486, and Q493. Orange indicates the subset of mutations with additional support as escape mutations from an independent (external) study. The replicated residues Y453 and F486 were also our top predictions of sites of escape mutations.

targets the spike-loop region near the edge of the ACE2 interface (Hansen et al., 2020). We trained neural networks in DMS antibody-escape data (Starr et al., 2021) on REGN10933. A map of the neural network-derived antibody escape phenotype for REGN10933 identified several peaks, that is, escape mutations, at K417, Y453, L455, F486, and Q493. Notably, we found substantial confirmation in an independent experimental study that had searched for escape mutations through deep sequencing of the passaged virus (Baum et al., 2020) (Figure 6D). The replicated residues Y453 and F486 were also the sites of our top escape mutation predictions.

## DISCUSSION

In this work, we leveraged neural networks to model the biochemical phenotypes of the spike RBD and the ACE2 PD. Mutations can directly influence protein conformational dynamics, stability, and associations/dissociations, that is, the sequence-structure-phenotype relationship (Ali and Vijayan, 2020). We examined the prediction performance of several neural network architectures designed to learn the features of the sequence-function landscape. Convolutional neural networks demonstrated superior performance over linear regression or multilayer perceptron in “decoding” the mutational information from sequence to phenotype. Furthermore, the sequence convolutional neural network, in general, outperformed the graph convolutional neural network (Tables S1 and S2), indicating that information learned from the sequence alone already guarantees substantial predictive performance in our investigated systems. The lower performance of the graph convolutional neural network might be related to the underlying network architecture as well as the physicochemical nature of the predicted phenotype, that is, the node average model (Fout et al., 2017) does not sufficiently capture relevant features from a structural perspective for binding affinity.

Our deep learning experiments confirmed the quality of the AAindex, which captures the intrinsic properties of each amino acid that are independent of the particular sequence, structure, or DMS effect scores. Our shuffling experiments showed that the AAindex significantly improved the prediction. Alternatively put, the use of a randomly generated AAindex was found to be sub-optimal, with the physicochemical properties encoded in the actual AAindex having a significantly greater impact on phenotype prediction. Nevertheless, we observed strong dependence of the biochemical phenotypes on the input sequence data alone, from which the convolutional neural networks could extract features and perform phenotype prediction reasonably well. As the protein sequence is the final product of gene transcription, the protein sequence—phenotype mapping from the neural networks has important consequences on bridging the gap between genetic variants and phenotype.

Investigating mutational effects is essential for deriving mechanistic insights and elucidating disease pathways. Molecular modeling tools such as molecular dynamics simulations and docking are frequently used toward these ends, with the clear benefit of obtaining atomic-detail understandings of biomolecular systems (Ali and Vijayan, 2020; Calcagnile et al., 2021). However, these tools are typically impeded by system complexity and computational demand, so that only a partial structure could be simulated for a limited number of mutational systems. As structural interactions comprise the central elements of molecular modeling, residues on the binding interface with high binding frequency or affinity are likely to be the focus, while dissociation effects like antibody escape are challenging to model. In fact, it has been reported that the rapid spread of COVID-19 has more to do with asymptomatic and pre-symptomatic transmission than enhanced receptor binding (Procko, 2020). In addition, allosteric effects are difficult to capture in molecular modeling yet critical to viral infection and evolution. It has been reported that the D614G variant that is located far from the RBD of the spike glycoprotein can facilitate ACE2 infection (Hou et al., 2020). Similarly, mutations near the ACE2’s chloride-binding sites, which are located far from the spike interface, may also alter the spike RBD binding (Heinzelman and Romero, 2021). Although achieving complete coverage of all mutations on the RBD-ACE2 complex, our models could be easily extended to other motifs with suitable data for the investigation of long-range dependencies such as allosteric effects. Notably, the models attained substantial predictive performance without modeling the long-range dependencies, confirming the centrality of the mutational landscape of the RBD-ACE2 complex. Some mutations could become prevalent as selection facilitates receptor binding. For instance, the N501Y mutation, which sits between two key binding residues, that is, T500 and G502 (Figures 5A–5C), has been reported to be responsible for the rapid spread of the virus in southeastern England, threatening the efficacies of the existing vaccines (Lauring and Hodcroft, 2021; Rambaut et al., 2020). We note that the models presented here could be readily repurposed to predict new relevant phenotypes within the same

framework. Our approach is an economical way to model high-dimensional data, including combinations of single-point mutations (see “data and code availability”), to gain insights into the therapeutic design (Figure 6).

It is worth noting that the tested neural networks achieved much better prediction performance on the spike RBD than on the ACE2 PD. The difference in prediction performance on the two interacting components of the protein complex could result from a difference in the data quality, that is, the DMS experiments had been performed in two separate studies. However, the difference may also be owing to a difference in the intrinsic properties of the virus and the receptor. Specifically, the human ACE2 phenotype may be less dependent on (or determined by) sequence mutation data. This finding was also supported by the fact that linear regression or fully connected networks failed to make a good prediction for the ACE2 PD.

### Limitations of the study

A limitation of our study was that it did not model long-range dependencies such as allosteric effects. Future work can also investigate the extent to which the integration of evolutionary information into neural network models improves performance. Finally, prediction may be further enhanced by leveraging network architectures not evaluated here.

### STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Datasets
  - Protein sequence encoding and structural description
  - Network architectures and model training
  - Joint model consisting of the neural network derived antibody-escape phenotypes
  - AAindex shuffling experiment
- QUANTIFICATION AND STATISTICAL ANALYSIS

### SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.isci.2022.104500>.

### ACKNOWLEDGMENTS

This research is supported by the National Institutes of Health (NIH) Awards R35HG010718, NIH/NIA AG068026, and R01HG011138.

### AUTHOR CONTRIBUTIONS

E.R.G. and B.W. designed the study and wrote the article. B.W. and E.R.G. performed the analyses. E.R.G. supervised and acquired funding for the study.

### DECLARATION OF INTERESTS

E.R.G. receives an honorarium from the journal *Circulation Research* of the American Heart Association, as a member of the Editorial Board.

Received: July 12, 2021

Revised: November 15, 2021

Accepted: May 26, 2022

Published: July 15, 2022

## REFERENCES

- Addetia, A., Crawford, K.H.D., Dingens, A., Zhu, H., Roychoudhury, P., Huang, M.-L., Jerome, K.R., Bloom, J.D., and Greninger, A.L. (2020). Neutralizing antibodies correlate with protection from SARS-CoV-2 in humans during a fishery vessel outbreak with a high attack rate. *J. Clin. Microbiol.* 58. e02107-20. <https://doi.org/10.1128/JCM.02107-20>.
- Ali, A., and Vijayan, R. (2020). Dynamics of the ACE2–SARS-CoV-2/SARS-CoV spike protein interface reveal unique mechanisms. *Sci. Rep.* 10, 14214. <https://doi.org/10.1038/s41598-020-71188-3>.
- Angermueller, C., Pärnamaa, T., Parts, L., and Stegle, O. (2016). Deep learning for computational biology. *Mol. Syst. Biol.* 12, 878. <https://doi.org/10.15252/msb.20156651>.
- Baum, A., Fulton, B.O., Wloga, E., Copin, R., Pascal, K.E., Russo, V., Giordano, S., Lanza, K., Negron, N., Ni, M., et al. (2020). Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual antibodies. *Science* 369, 1014–1018. <https://doi.org/10.1126/science.abd0831>.
- Blanco, J.D., Hernandez-Alias, X., Cianferoni, D., and Serrano, L. (2020). In silico mutagenesis of human ACE2 with S protein and translational efficiency explain SARS-CoV-2 infectivity in different species. *PLoS Comput. Biol.* 16, e1008450. <https://doi.org/10.1371/journal.pcbi.1008450>.
- Calcagnile, M., Forgez, P., Iannelli, A., Bucci, C., Alifano, M., and Alifano, P. (2021). Molecular docking simulation reveals ACE2 polymorphisms that may increase the affinity of ACE2 with the SARS-CoV-2 Spike protein. *Biochimie* 180, 143–148. <https://doi.org/10.1016/j.biochi.2020.11.004>.
- Chan, K.K., Dorosky, D., Sharma, P., Abbasi, S.A., Dye, J.M., Kranz, D.M., Herbert, A.S., and Procko, E. (2020). Engineering human ACE2 to optimize binding to the spike protein of SARS coronavirus 2. *Science* 369, 1261–1265. <https://doi.org/10.1126/science.abc0870>.
- Elbe, S., and Buckland-Merrett, G. (2017). Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob. Chall.* 1, 33–46. <https://doi.org/10.1002/gch2.1018>.
- Fout, A., Byrd, J., Shariat, B., and Ben-Hur, A. (2017). Protein interface prediction using graph convolutional networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, U. von Luxburg and I. Guyon, eds. (Curran Associates Inc), pp. 6533–6542.
- Fowler, D.M., Araya, C.L., Fleishman, S.J., Kellogg, E.H., Stephany, J.J., Baker, D., and Fields, S. (2010). High-resolution mapping of protein sequence-function relationships. *Nat. Methods* 7, 741–746. <https://doi.org/10.1038/nmeth.1492>.
- Fowler, D.M., and Fields, S. (2014). Deep mutational scanning: a new style of protein science. *Nat. Methods* 11, 801–807. <https://doi.org/10.1038/nmeth.3027>.
- Gelman, S., Romero, P.A., and Gitter, A. (2020). Neural networks to learn protein sequence-function relationships from deep mutational scanning data. Preprint at bioRxiv. <https://doi.org/10.1101/2020.10.25.353946>.
- Gray, V.E., Hause, R.J., Luebeck, J., Shendure, J., and Fowler, D.M. (2018). Quantitative missense variant effect prediction using large-scale mutagenesis data. *Cell Syst.* 6, 116–124.e3. <https://doi.org/10.1016/j.cels.2017.11.003>.
- Greaney, A.J., Starr, T.N., Gilchuk, P., Zost, S.J., Binshtein, E., Loes, A.N., Hilton, S.K., Huddleston, J., Eguia, R., Crawford, K.H.D., et al. (2021). Complete mapping of mutations to the SARS-CoV-2 spike receptor-binding domain that escape antibody recognition. *Cell Host Microbe* 29, 44–57.e9. <https://doi.org/10.1016/j.chom.2020.11.007>.
- Han, Y., and Král, P. (2020). Computational design of ACE2-based peptide inhibitors of SARS-CoV-2. *ACS Nano* 14, 5143–5147. <https://doi.org/10.1021/acsnano.0c02857>.
- Hansen, J., Baum, A., Pascal, K.E., Russo, V., Giordano, S., Wloga, E., Fulton, B.O., Yan, Y., Koon, K., Patel, K., et al. (2020). Studies in humanized mice and convalescent humans yield a SARS-CoV-2 antibody cocktail. *Science* 369, 1010–1014. <https://doi.org/10.1126/science.abd0827>.
- Heinzelman, P., and Romero, P.A. (2021). Discovery of human ACE2 variants with altered recognition by the SARS-CoV-2 spike protein. *PLoS One* 16, e0251585. <https://doi.org/10.1371/journal.pone.0251585>.
- Hou, Y.J., Chiba, S., Halfmann, P., Ehre, C., Kuroda, M., Dinnon, K.H., Leist, S.R., Schäfer, A., Nakajima, N., Takahashi, K., et al. (2020). SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo. *Science* 370, 1464–1468. <https://doi.org/10.1126/science.abe8499>.
- Hussain, M., Jabeen, N., Raza, F., Shabbir, S., Baig, A.A., Amanullah, A., and Aziz, B. (2020). Structural variations in human ACE2 may influence its binding with SARS-CoV-2 spike protein. *J. Med. Virol.* 92, 1580–1586. <https://doi.org/10.1002/jmv.25832>.
- Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., and Kanehisa, M. (2008). AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.* 36, D202–D205. <https://doi.org/10.1093/nar/gkm998>.
- Lan, J., Ge, J., Yu, J., Shan, S., Zhou, H., Fan, S., Zhang, Q., Shi, X., Wang, Q., Zhang, L., and Wang, X. (2020). Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* 581, 215–220. <https://doi.org/10.1038/s41586-020-2180-5>.
- Lauring, A.S., and Hodcroft, E.B. (2021). Genetic variants of SARS-CoV-2—what do they mean? *JAMA* 325, 529. <https://doi.org/10.1001/jama.2020.27124>.
- Leek, J.T., and Storey, J.D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* 3, e161. <https://doi.org/10.1371/journal.pgen.0030161>.
- Li, B., Yang, Y.T., Capra, J.A., and Gerstein, M.B. (2020). Predicting changes in protein thermodynamic stability upon point mutation with deep 3D convolutional neural networks. *PLoS Comput. Biol.* 16, e1008291. <https://doi.org/10.1371/journal.pcbi.1008291>.
- Listgarten, J., Kadie, C., Schadt, E.E., and Heckerman, D. (2010). Correction for hidden confounders in the genetic analysis of gene expression. *Proc. Natl. Acad. Sci. U S A* 107, 16465–16470. <https://doi.org/10.1073/pnas.1002425107>.
- McInnes, L., Healy, J., and Melville, J. (2020). UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at arXiv. <https://doi.org/10.48550/arXiv.1802.03426>.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.
- Pietzner, M., Wheeler, E., Carrasco-Zanini, J., Raffler, J., Kerrison, N.D., Oerton, E., Auyeung, V.P.W., Luan, J., Finan, C., Casas, J.P., et al. (2020). Genetic architecture of host proteins involved in SARS-CoV-2 infection. *Nat. Commun.* 11, 6397. <https://doi.org/10.1038/s41467-020-19996-z>.
- Rambaut, A., Loman, N., Pybus, O., Barclay, W., Barrett, J., Carabelli, A., Connor, T., Peacock, T., Robertson, D., and Volz, E. (2020). Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. *Virological*. <https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563>.
- Procko, E. (2020). Deep mutagenesis in the study of COVID-19: a technical overview for the proteomics community. *Expert Rev. Proteomics* 17, 633–638. <https://doi.org/10.1080/14789450.2020.1833721>.
- Sarkisyan, K.S., Bolotin, D.A., Meer, M.V., Usmanova, D.R., Mishin, A.S., Sharonov, G.V., Ivankov, D.N., Bozhanova, N.G., Baranov, M.S., Soyomez, O., et al. (2016). Local fitness landscape of the green fluorescent protein. *Nature* 533, 397–401. <https://doi.org/10.1038/nature17995>.
- Shu, Y., and McCauley, J. (2017). GISAID: global initiative on sharing all influenza data – from vision to reality. *Euro Surveill.* 22, 30494. <https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494>.
- Song, H., Bremer, B.J., Hinds, E.C., Raskutti, G., and Romero, P.A. (2021). Inferring protein sequence-function relationships with large-scale positive-unlabeled learning. *Cell Syst.* 12, 92–101.e8. <https://doi.org/10.1016/j.cels.2020.10.007>.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014).

Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15, 1929–1958.

Starr, T.N., Greaney, A.J., Addetia, A., Hannon, W.W., Choudhary, M.C., Dingens, A.S., Li, J.Z., and Bloom, J.D. (2021). Prospective mapping of viral mutations that escape antibodies used to treat COVID-19. *Science* 371, 850–854. <https://doi.org/10.1126/science.abf9302>.

Starr, T.N., Greaney, A.J., Hilton, S.K., Ellis, D., Crawford, K.H.D., Dingens, A.S., Navarro, M.J., Bowen, J.E., Tortorici, M.A., Walls, A.C., et al. (2020). Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on

folding and ACE2 binding. *Cell* 182, 1295–1310.e20. <https://doi.org/10.1016/j.cell.2020.08.012>.

Stein, A., Fowler, D.M., Hartmann-Petersen, R., and Lindorff-Larsen, K. (2019). Biophysical and mechanistic models for disease-causing protein variants. *Trends Biochem. Sci.* 44, 575–588. <https://doi.org/10.1016/j.tibs.2019.01.003>.

Stevens, S.S. (1946). On the theory of scales of measurement. *Science* 103, 677–680. <https://doi.org/10.1126/science.103.2684.677>.

Warner, S. (1958). The topology of compact convergence on continuous function spaces.

*Duke Math. J.* 25, 265–282. <https://doi.org/10.1215/S0012-7094-58-02523-7>.

Zhang, H., Penninger, J.M., Li, Y., Zhong, N., and Slutsky, A.S. (2020). Angiotensin-converting enzyme 2 (ACE2) as a SARS-CoV-2 receptor: molecular mechanisms and potential therapeutic target. *Intensive Care Med.* 46, 586–590. <https://doi.org/10.1007/s00134-020-05985-9>.

Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., Si, H.-R., Zhu, Y., Li, B., Huang, C.-L., et al. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 579, 270–273. <https://doi.org/10.1038/s41586-020-2012-7>.

## STAR★METHODS

## KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Other		
Spike RBD	Starr et al.	N/A
ACE2 PD	Heinzelman et al.	N/A
Neutralizing antibodies	Greaney et al.	N/A
Monoclonal antibody	Starr et al.	N/A
Software and algorithms		
Neural network engine (nn4dms)	Gelman et al.	N/A
Node average model (pipgcn)	Fout et al.	N/A

## RESOURCE AVAILABILITY

## Lead contact

Further information and requests for resources should be directed to the Lead Contact, Eric R. Gamazon ([eric.gamazon@vumc.org](mailto:eric.gamazon@vumc.org)).

## Materials availability

This study did not generate any new unique reagents.

## Data and code availability

- The data and results are available in the [supplemental information](#).
- The code for this project (as well as the full set of trained models and a sample application for predicting the effects of combinations of single-point mutations) is available from GitHub (<https://github.com/gamazonlab/DeepMutScan>).
- Any additional information is available from the [lead contact](#) upon request.

## METHOD DETAILS

## Datasets

We performed our deep learning experiments on publicly available spike RBD and ACE2 PD DMS datasets ([Greaney et al., 2020](#); [Heinzelman and Romero, 2021](#); [Starr et al., 2020](#)), where a 6:2:2 ratio was implemented for random training/tuning/test split. Binding affinity (i.e.,  $\Delta\log_{10}K_{D,app}$ , where  $K_{D,app}$  is the 'apparent dissociation constant' in molar units) and protein expression (i.e., quantified as the difference in log-mean fluorescence intensity (MFI) relative to wild-type:  $\Delta\log(\text{MFI}) = \log(\text{MFI}_{\text{variant}}) - \log(\text{MFI}_{\text{wild-type}})$ ) of the spike RBD as well as the relative binding score (a "site importance" index quantifying how an amino acid substitution affects a protein's binding ([Song et al., 2021](#))) of the ACE2 PD were considered. We also leveraged a deep mutational scan of the RBD to determine how variants from spike RBD responded to antibody binding from 10 human monoclonal antibodies, of which nine target SARS-CoV-2 and one targets SARS-CoV. The original mutational effect scores - "escape fraction" - have been normalized via Box Cox transformation.

Models on the avGFP protein (an independent DMS dataset) were also retrained and tested in the RBD expression dataset ([Gelman et al., 2020](#); [Sarkisyan et al., 2016](#)). For this purpose, the wild-type sequence for avGFP was reduced so that the overall sequence length matched that of the spike RBD. Accordingly, the original DMS dataset for avGFP was modified to only include the variant-score profile of the sliced sequence.



### Protein sequence encoding and structural description

We curated the original AAindex database, a resource of 566 physicochemical and biochemical properties (e.g., polarizability parameter, residue volume, solvation free energy, and other attributes) for each of the 20 amino acids (Kawashima et al., 2008). The data underwent dimensionality reduction to a 20\*19 matrix (reducing the original 566 [correlated] properties to 19 orthogonal Principal Components), which was loaded to the neural networks as input features (Gelman et al., 2020). The one-hot encoding converted categorical data to numerical form (Stevens, 1946). Specifically, for a given sequence, amino acids from 20 different “categories”, were encoded into binary variables where “1” and “0” denote the presence and non-presence, respectively, of an amino acid at a position. The wild-type and mutated protein sequence were encoded using both one-hot encoding and AAindex.

### Network architectures and model training

Here, we extend a supervised learning framework for application to interrogating the biochemical phenotypes relevant to the virulence of SARS-CoV-2 (Gelman et al., 2020). We utilized different neural network architectures including linear regression, multilayer perceptron (fully connected), sequence convolutional neural network, and graph convolutional neural network. These architectures can be viewed as directed computational graphs organized as nodes in a series of hidden layers  $h_{(k)}$

$$h_{(k)} = \varphi(\mathbf{W}_{(k)}h_{(k-1)} + b_{(k)})$$

with a matrix of weights  $\mathbf{W}_{(k)}$  and biases  $b_{(k)}$ . The network architectures are implemented as compositions of functions (Equation 1), reflecting the universal approximation properties of the architectures with respect to the compact convergence (compact-open) topology (Warner, 1958).

Convolutional neural networks perform convolution, activation, pooling, and flattening to learn features from input data. Briefly, convolution is a mathematical operation involving a convolution operator  $m$  and input  $s$ :

$$(m * s)(t) = \int_{-\infty}^{\infty} m(x) \cdot s(t - x) dx$$

This one-dimensional definition can be extended to arbitrary dimensions. This formulation has a discrete version:

$$(m * s)(t) = \sum_{x=1}^N m[x] \cdot s[t + x - 1]$$

The convolution operation can be viewed as a weight assignment scheme. Different operators or *filters* can be applied to the input data to derive an optimally predictive set of features. Activation introduces non-linearity and pooling helps to reduce overfitting. A fully-connected neural network is then applied to the resulting feature vector (after flattening) to obtain the final prediction.

Minimization of the empirical loss across the training samples

$$\hat{g} = \operatorname{argmin}_g \frac{1}{N} \sum_{i=1}^N L(y_i, g(x_i))$$

yields an estimate of the mutation- or genetically- determined  $\hat{g}(x)$ , whose properties can be explored (and compared with the original  $g$ ) and which may enable causal inference on a biochemical phenotype. We trained the models using mean squared error loss and the Adam optimizer with dropout (Srivastava et al., n.d.) as the regularization method. To prevent overfitting, the dropout method randomly drops a proportion of the nodes (in our case, dropout rate of 20%) and their connections from the neural network during the training. For activation function  $\varphi$  (Equation 1), a leaky ReLU was applied, which permits a small, positive gradient for a non-active unit

$$\varphi(x) = \begin{cases} x, & x > 0 \\ 0.01x, & x \leq 0 \end{cases}$$

Neural networks were implemented using TensorFlow. We set the maximum possible number of epochs to 400 with early stopping.

The uncertainty associated with the neural network model can be quantified. Let  $\omega$  be the set of all model parameters, i.e., the weights  $\mathbf{W}_{(k)}$  and biases  $b_{(k)}$ . Then the posterior probability  $P(\omega | X_{\text{train}}, Y_{\text{train}})$  of the set

of parameters conditional on the training sequence data  $X_{\text{train}}$  and the training output data  $Y_{\text{train}}$  is, according to Bayes' rule, given by

$$P(\omega|X_{\text{train}}, Y_{\text{train}}) = P(Y_{\text{train}}|X_{\text{train}}, \omega)P(\omega) / \int P(Y_{\text{train}}|X_{\text{train}}, \omega')P(\omega')d\omega'$$

Here,  $P(\omega)$  is the *prior* distribution of the set of parameters and  $P(Y_{\text{train}}|X_{\text{train}}, \omega)$  may be assumed to follow

$$P(Y|X, \omega) = \begin{cases} N(\widehat{g}(x), V), & \text{continuous } Y \\ \exp(\widehat{g}_a(x)) / \sum_{a'} \exp(\widehat{g}_{a'}(x)), & Y = a \end{cases}$$

That is, in the continuous case, the likelihood is Gaussian with mean  $\widehat{g}(x)$  and covariance matrix  $V$  whereas for classification, the conditional probability is a softmax likelihood (assuming the output layer is a softmax layer and the cross-entropy is the loss function). The posterior probability  $P(\omega | X_{\text{train}}, Y_{\text{train}})$  can be calculated using Markov Chain Monte Carlo (MCMC), as the integral (i.e., the so-called *model evidence*  $P(Y_{\text{train}}|X_{\text{train}})$ ) is known to be typically intractable and have no analytical formulation. Prediction of the phenotype  $y$  for any given input  $x$  can then be quantified through the following inference

$$P(y|x, X_{\text{train}}, Y_{\text{train}}) = \int P(y|x, \omega')P(\omega'|X_{\text{train}}, Y_{\text{train}})d\omega'$$

Note the left hand side of this equation can also be sampled using the neural network.

### Joint model consisting of the neural network derived antibody-escape phenotypes

The spike RBD binding affinity was jointly modeled using the convolutional neural network derived (i.e., estimated, mutation-mediated) antibody-escape phenotypes for the ten Abs. The ten "escape fraction" features were scaled and randomly split into 80% training and 20% testing subsets for the Random Forest regressor implemented using scikit-learn (Pedregosa et al., 2011). We applied hyper parameter tuning to ensure the best performance.

### AAindex shuffling experiment

We performed a shuffling experiment on the AAindex to assess the extent to which the selected amino acid physicochemical properties impact the performance of the neural network architectures. All elements of the AAindex matrix were re-sampled in-place so that the original covariances among rows or columns were broken. As a result, we generated 100 new (pseudo) AAindex tables. The p-value, i.e., the probability, if the null hypothesis  $H_0$  (of no gain) is true, that the prediction performance  $\rho$  (in the test set) is at least as great as the value  $\rho_{\text{actual}}$  from the actual AAindex is calculated as follows

$$p : = P(\rho \geq \rho_{\text{actual}} | H_0) \approx (|\rho_{\text{sim}} \geq \rho_{\text{actual}}| + 1) / (n_{\text{sim}} + 1)$$

where  $\rho_{\text{sim}}$  is the Spearman correlation from a shuffled dataset,  $n_{\text{sim}}$  is the total number of shuffled datasets, and  $||$  is the count operator. Using each pseudo AAindex, the model training, including dropout regularization, was performed in the same exact way as for the actual AAindex table.

## QUANTIFICATION AND STATISTICAL ANALYSIS

Correlation analysis was performed in R (3.6.1). Predictive performance of neural networks on ACE2 binding affinity ( $p < 2.2 \times 10^{-16}$ ; Figure 2) and on spike glycoprotein expression ( $p < 2.2 \times 10^{-16}$ ; Figure 3) were determined using the Spearman correlation between observed and predicted values in an independent test set. P-values for the following comparisons were determined using the Mann-Whitney U test: 1) the neural network derived binding affinity with the ACE2 receptor between key binding residues and the remaining residues in the spike RBD ( $p = 5.19 \times 10^{-18}$ ); 2) the difference between the two residue classes for the original binding affinity phenotype ( $p = 6.9 \times 10^{-14}$ ); 3) the neural network derived binding affinity between sites of escape mutations of the antibody COV2-2499 and the remaining key binding residues in the spike RBD ( $p = 3.2 \times 10^{-4}$ ); 4) the neural network derived RBD protein expression profile between the two residues classes in the previous comparison ( $p = 1.03 \times 10^{-19}$ ); and 5) the neural network derived binding of ACE2 between the key binding residues in the ACE2 receptor (paired with those of the spike RBD) and the remaining residues ( $p = 2.39 \times 10^{-7}$ ). Plots were generated using ggplot2 (3.3.1).